Emerging Entity Discovery Using Web Sources

Lei Zhang[1], Tianxing Wu[2], Liang Xu[3], Meng Wang[3], Guilin Qi[3], and
Harald Sack[1]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[2] Nanyang Technological University, Singapore
[3] Southeast University, China
{lei.zhang,harald.sack}@fiz-karlsruhe.de
{wutianxing}@ntu.edu.sg
{lxu,mwang,gqi}@seu.edu.cn

Abstract. The rapidly increasing amount of entities in knowledge bases (KBs) can be beneficial for many applications, where the key issue is to link entity mentions in text with entities in the KB, also called *entity linking* (EL). Many methods have been proposed to tackle this problem. However, the KB can never be complete, such that *emerging entity discovery* (EED) is essential for detecting emerging entities (EEs) that are mentioned in text but not yet contained in the KB. In this paper, we propose a new topic-driven approach to EED by representing EEs using the context harvested from online Web sources. Experimental results show that our solution outperforms the state-of-the-art methods in terms of F1 measure for the EED task as well as Micro Accuracy and Macro Accuracy in the full EL setting.

1 Introduction

As large knowledge bases (KBs) of individual entities became available, it enabled the linking of words or phrases in text to entities in the KB. The challenges of entity linking (EL) lie in entity recognition and disambiguation. The first stage, i.e., entity recognition (ER), is to identify the word sequences in text that refer to an entity, also called *mentions*, for which no KB is required. The second stage, i.e., entity disambiguation (ED), aims at mapping ambiguous mentions onto entities like persons, organizations or movies in the KB.

In spite of the rapidly increasing quantities of entities in the KB, the knowledge can never be complete due to (1) the ever-changing world, e.g., *new entities* appear under the same names as existing ones in the KB, and (2) a *long-tail of entities* that are not captured by the KB because they lack the importance. In [1], a survey to thoroughly investigate various types of challenges that arise from out-of-KB entities in the context of EL has been provided. We refer to such out-of-KB entities as *emerging entities* (EEs) and EL methods must cope with this issue, i.e., mentions that have no corresponding entities in the KB. In this work, the key problem is to determine when a mention refers to an EE by discriminating it against the existing candidate entities in the KB. The task is also called *emerging entity discovery* (EED). The examples of two kinds of EEs, i.e., new entities and long-tail entities, are given in the following.

Example (New Entities) Suppose an EL method is fed with the input text "*Alphabet*, Google's new parent company, is boldly restructuring the search engine giant and its subsidiaries." from one of the early news articles on this topic, before the entity Alphabet_Inc. being added into the KB due to its lagging behind news [2]. The EL method needs to determine that the mention "*Alphabet*" does not refer to Alphabet_(poetry_collection), a 1981 book by Danish poet Inger Christensen that exists in the KB, e.g., Wikipedia, for quite a long time, and instead should be mapped to an EE.

Example (Long-tail Entities) Consider the news about *Michael Jordan*, a professor of English at the University of St. Thomas, who does not exist in the KB. An EL method needs to decide that the mention "*Michael Jordan*" in such a news should refer to an EE, instead of a candidate entity in the KB, such as Michael_Jordan, an American retired professional basketball player, or Michael_I._Jordan, a professor in machine learning, statistics, and artificial intelligence at the University of California, Berkeley.

However, most existing EL methods cannot robustly deal with mentions that have no correct entity candidate in the KB. As soon as there is a candidate entity for a mention in the input text, these algorithms are destined to choose one. In order to identify mentions that have no good match in the KB, a simple solution is to employ a confidence threshold to disregard the candidate entities in the KB yielded by an algorithm, such that a mention with a low confidence score for all KB entities is determined to refer to an EE.

In contrast, Hoffart et al. [3] has introduced a new approach, which models an EE as a set of weighted keyphrases collected from news articles by looking back some days before the publishing date of the input text and introduces an additional EE candidate for each mention. Once the candidate space is expanded with EEs, the EL problem is fed back to a prior EL method (i.e., AIDA [4]), which is based on the same keyphrase features, such that it can treat EEs in the same way as it treats KB entities. This is the state-of-the-art method for EED and also the most related work to ours.

The main drawback of the method in [3] is that adding an EE candidate for each mention would potentially introduce noise, which, as showed in our experiments, resulted in degraded EL decisions for the mentions referring to an existing entity in the KB. In order to address this problem, we propose a new solution to EED. Different from [3], our approach employs a prior EL method as a black box and takes its results (i.e., the mappings between each mention and its most likely referent entity in the KB) as the input for further EE detection. In addition, it does not affect existing EL decisions for KB entities yielded by the prior EL method.

Towards a robust solution to EED in the context of EL, we provide in this work the following contributions:

- In order to capture both new entities and long-tail entities, we accurately harvest the context of such emerging entities from online Web sources using a Web search engine.

- We enrich KB entities and EE candidates with an appropriate representation as topic distributions of their contexts, based on that develop a principled method of topic-driven EED.
- The experiments conducted on a benchmark dataset for EED show the superior quality of our solution compared to the state-of-the-art methods in terms of F1 measure of EE results as well as Micro Accuracy and Macro Accuracy of EL results.

2 Approach

Firstly, we formally formulate the task of EL by taking into account EEs. Then, we present our solution to EED in the context of EL.

Definition 1 (Entity Linking). *Let $M = \{m_1, \ldots, m_k\}$ denote the set of all words and phrases in a document D. Given a knowledge base KB containing a set of entities $E = \{e_1, \ldots, e_n\}$, the objective of* entity linking (ER) *is to determine the referent entities for the mentions in M, where two functions are to be found: (1) an* entity recognition (ER) *function $f : D \rightarrow 2^M$ that aims to identify the set of entity mentions $\mu \subseteq M$ from D, and (2) an* entity disambiguation (ED) *function $g : \mu \rightarrow E \cup \{EE\}$ that maps the set of mentions μ yielded by the recognition function to entities in KB or to* emerging entities *that are not yet contained in KB, denoted by the label EE.*

We assume that the KB used in this work is Wikipedia, or any others where each entity has a corresponding Wikipedia page, such as DBpedia [5] and YAGO [6]. Now we recap the computational model of EL. Firstly, the text document is processed by a method for ER, e.g., the Stanford NER Tagger [7], which detects the boundaries of entity mentions. These detected mentions serve as the input of ED in the next step, where the goal is to infer the actual referent entities in the KB or the label EE in case that the corresponding entities do not exist in the KB. In many existing EL methods (e.g., [8–10]), the output also includes a confidence score for each mapping between a mention and its most likely referent entity in the KB.

In our approach, we firstly employ a probabilistic EL method [10], which models the interdependence between different EL decisions as a graph to capture both local mention-entity compatibility and global entity-entity coherence, where evidences for EL can be collectively reinforced into high-confidence decisions based on a random walk process. In principle, many EL methods can be applied here as long as they provide a confidence score for the individual outputs (i.e., mention-entity mappings). Instead of thresholding on the confidence score to directly determine EE, we only use a threshold to filter out the mentions that have a high-confidence mapping to an existing entity in the KB. Then, the remaining mentions are considered as EE candidates and fed into an additional model of EED, which involves *entity context harvesting* (Sec. 2.1), *context representation learning* (Sec. 2.2) and *EE detection* (Sec. 2.3).

2.1 Entity Context Harvesting

For each mention m as an EE candidate, we firstly collect its entity context from Wikipedia, where each page describes a corresponding KB entity. Also, a Wikipedia page often contains hyperlinks pointing to the pages of other entities and the anchor text of a hyperlink provides the mention of the linked entity. Based on that, we define the context of m w.r.t. KB entities, denoted by $\mathbf{C_{KB}} = \{p_i\}_{i=1}^{|\mathbf{C_{KB}}|}$, as a set of Wikipedia pages p_i linked from the anchor text m, where each page corresponds to a KB entity referred to by m.

Although EEs do not have textual information in Wikipedia, there might exist some associated Web pages. Therefore, we decide to acquire the entity context for a mention m as an EE candidate by querying the Web with a search engine[4]. To accurately get such context, we firstly perform coreference resolution [11] to find all expressions that refer to the same entity as m in the input document and based on POS tagging [12] to extract the noun phrases that co-occur with all coreferences of m in the same sentences. Then, the mention m and the extracted noun phrases are jointly submitted to the search engine to retrieve a set of relevant Web pages p_j as the *actual* entity context of m, denoted by $\mathbf{C_{Actual}} = \{p_j\}_{j=1}^{|\mathbf{C_{Actual}}|}$.

Given a mention m, its actual entity context $\mathbf{C_{Actual}}$ could correspond to either a KB entity or an EE, while $\mathbf{C_{KB}}$ captures the context of all existing entities in the KB that can be referred to by m. In order to perform EED on m as an EE candidate, our basic idea is to check if the actual entity context $\mathbf{C_{Actual}}$ is dissimilar enough to the KB entity context $\mathbf{C_{KB}}$. If so, we assume that there should be an EE that has quite different context from all the referent KB entities of m. To compare the textual contexts $\mathbf{C_{KB}}$ and $\mathbf{C_{Actual}}$, the bag-of-words (BOW) model is the most common method to represent text as vectors, and based on that we can apply standard functions (e.g., Euclidian distance, dot product and cosine) to calculate the vector similarity. However, the textual contexts are extracted from different sources, i.e., Wikipedia and various websites, that vary a lot in wording styles, such that the same words in $\mathbf{C_{KB}}$ and $\mathbf{C_{Actual}}$ could be of low frequency even though they share common information. Therefore, the BOW model may not work well in this scenario.

2.2 Context Representation Learning

To address the problem of the BOW model, we try to discover the topics of $\mathbf{C_{KB}}$ and $\mathbf{C_{Actual}}$ with a topic model, i.e., Latent Dirichlet allocation (LDA) [13], such that we can compare these two kinds of contexts based on their representations as topic distributions.

Suppose the corpus $\mathbf{C} = \mathbf{C_{KB}} \cup \mathbf{C_{Actual}} = \{p_j\}_{j=1}^{|\mathbf{C_{KB}}|} \cup \{p_i\}_{i=1}^{|\mathbf{C_{Actual}}|}$ contains $|\mathbf{C_{KB}}| + |\mathbf{C_{Actual}}|$ documents, W distinct words and K topics expressed over the individual words in these documents. The topic indicator variable is denoted by $z_{in} \in [1, K]$ and $z_{jn} \in [1, K]$ for the n-th word in the Wikipedia page

[4] We choose Microsoft Bing as the Web search engine in this work.

Algorithm 1: Generative Process of $\mathbf{C_{KB}}$ and $\mathbf{C_{Actual}}$

```
initialize: (1) set the number of topics $K$;
            (2) set the values of Dirichlet priors $\alpha$ and $\beta$;
foreach topic $k \in [1, K]$ do
    sample: $\phi_k \sim Dir(\beta)$;
sample: $\theta \sim Dir(\alpha)$;
foreach Wikipedia page $p_i \in \mathbf{C_{KB}}$ do
    foreach of $N_i$ word $w_{in} \in p_i$ do
        sample: $z_{in} \sim Multinonimal(\theta_i)$;
        sample: $w_{in} \sim Multinonimal(\phi_{z_{in}})$;
foreach Web page $p_j \in \mathbf{C_{Actual}}$ do
    foreach of $N_j$ word $w_{jn} \in p_j$ do
        sample: $z_{jn} \sim Multinonimal(\theta_j)$;
        sample: $w_{jn} \sim Multinonimal(\phi_{z_{jn}})$;
```

$p_i \in \mathbf{C_{KB}}$ and in the Web page $p_j \in \mathbf{C_{Actual}}$, respectively. For each topic k, the corresponding word distribution is represented by a W-dimensional multinomial distribution ϕ_k with entry $\phi_k^w = P(w|z = k)$. In addition, we employ a K-dimensional multinomial distribution $\theta_i = \{\theta_i^k\}_{k=1}^K$ and $\theta_j = \{\theta_j^k\}_{k=1}^K$ with $\theta_i^k = \theta_j^k = P(z = k)$ to describe the topic distributions of each $p_i \in \mathbf{C_{KB}}$ and each $p_j \in \mathbf{C_{Actual}}$. Following the convention of LDA, the hyperparameters α and β are set as the Dirichlet priors. Based on that, the generative process of $\mathbf{C_{KB}}$ and $\mathbf{C_{Actual}}$ is described in Algorithm 1. Accordingly, the probability of generating both $\mathbf{C_{KB}}$ and $\mathbf{C_{Actual}}$ can be expressed as follows:

$$
\begin{aligned}
& P(\mathbf{C_{KB}}, \mathbf{C_{Actual}}|\alpha, \beta) \\
= & \prod_{k=1}^{K} P(\phi_k|\beta) \prod_{i=1}^{|\mathbf{C_{KB}}|} \Big[P(\theta_i|\alpha) \big[\prod_{n=1}^{N_i} \sum_{z_{in}} P(z_{in}|\theta_i) P(w_{in}|z_{in}, \phi) \big] \Big] \\
\times & \prod_{j=1}^{|\mathbf{C_{Actual}}|} \Big[P(\theta_j|\alpha) \big[(\prod_{n=1}^{N_j} \sum_{z_{jn}} P(z_{jn}|\theta_j) P(w_{jn}|z_{jn}, \phi) \big] \Big]
\end{aligned}
\tag{1}
$$

It is usually intractable to perform exact inference in such a probabilistic model, therefore we adopt Gibbs sampling [14] to conduct approximate inference. More specifically, we estimate the posterior distribution on z_{in} (z_{jn}) and then sample the topic for each word w_{in} (w_{jn}). Based on the sampling results after a sufficient number of iterations, we can estimate the parameters θ_i and θ_j that represent the topic distributions of each Wikipedia page $p_i \in \mathbf{C_{KB}}$ and each Web page $p_j \in \mathbf{C_{Actual}}$. In our experiments, we set the number of topics K as 25. For the hyperparameters α and β, we take the fixed values, i.e., $\alpha = 50/K$, $\beta = 0.01$.

2.3 Emerging Entity Detection

Given the topics derived from $\mathbf{C_{KB}}$ and $\mathbf{C_{Actual}}$, we represent the topic distributions of the KB entity context, denoted by θ^{KB} and the actual entity context, denoted by θ^{Actual}, as follows:

$$\theta^{KB} = \frac{1}{|\mathbf{C_{KB}}|} \sum_{i=1}^{|\mathbf{C_{KB}}|} \theta_i \tag{2}$$

$$\theta^{Actual} = \frac{1}{|\mathbf{C_{Actual}}|} \sum_{j=1}^{|\mathbf{C_{Actual}}|} \theta_j \tag{3}$$

Then, we measure the difference between $\mathbf{C_{KB}}$ and $\mathbf{C_{Actual}}$ using the Kullback Leibler (KL) divergence between the topic distributions θ^{KB} and θ^{Actual} as follows:

$$D_{KL}(\theta^{KB}||\theta^{Actual}) = \sum_{t=1}^{K} \theta_t^{KB} \cdot \log_2 \frac{\theta_t^{KB}}{\theta_t^{Actual}} \tag{4}$$

Eq. 4 measures how one probability distribution, i.e., θ^{KB}, diverges from another distribution, i.e., θ^{Actual}, which is equal to zero when $\theta_t^{KB} = \theta_t^{Actual}$ for all topics t. As the KL divergence is asymmetric, we apply a symmetric measure to calculate the final divergence between $\mathbf{C_{KB}}$ and $\mathbf{C_{Actual}}$ as follows:

$$D(\mathbf{C_{KB}}, \mathbf{C_{Actual}}) = \frac{1}{2}[D_{KL}(\theta^{KB}||\theta^{Actual}) + D_{KL}(\theta^{Actual}||\theta^{KB})] \tag{5}$$

Based on Eq. 5, we learn a threshold τ for $D(\mathbf{C_{KB}}, \mathbf{C_{Actual}})$ to determine whether a mention m refers to a KB entity or an EE. The assumption behind it is that if the actual entity is an EE that is not yet contained in the KB, its context $\mathbf{C_{Actual}}$ should be generated by a topic distribution that is to some extent divergent from the topic distribution of the context $\mathbf{C_{KB}}$ for the candidate KB entities.

3 Experiments and Achieved Results

We now discuss the experiments we have conducted to assess the performance of our approach to EED.

3.1 Experimental Settings

We firstly describe the experimental settings with respect to *Data* and *Evaluation Measures*.

Data. In the experiments, we employ the AIDA-EE dataset, also used by [3], which consists of 150 news articles published on October 1st and 150 published on November 1st, 2010, taken from the GigaWord 5 corpus [15], where each

Total number of documents	300
Total number of mentions	9,976
Total number of mentions with EE	561
Average number of words per article	538
Average number of mentions per article	33
Average number of entities per mention	104

Table 1. AIDA-EE GigaWord dataset statistics.

mention was manually annotated with EE if the referent entity is not present in Wikipedia as of 2010-08-17, otherwise the correct entity. The statistics of the dataset is given in Table 1. Accordingly, the knowledge base used in the experiments is based on the Wikipedia snapshot from 2010-08-17.

Evaluation Measures. We evaluate the quality of the overall EL (for both KB entities and EEs) with *Micro Accuracy* and *Macro Accuracy*. Additional measures to evaluate the quality of EED include *EE Precision*, *EE Recall* and *EE F1*. Let D be the collection of documents, G_d be all mentions in document $d \in D$ annotated by a human annotator with a gold standard entity, G_d^{EE} be the subset of G_d annotated with an emerging entity EE, A_d be all mentions in $d \in D$ automatically annotated by a method and A_d^{EE} be the subset of A_d annotated with EE. Based on that, the measures of *Micro Accuracy* and *Macro Accuracy* are defined as follows:

$$\text{Micro Accuracy} = \frac{|\bigcup_{d \in D} G_d \cap \bigcup_{d \in D} A_d|}{|\bigcup_{d \in D} G_d|} \tag{6}$$

$$\text{Macro Accuracy} = \frac{\sum_{d \in D} \frac{|G_d \cap A_d|}{|G_d|}}{|D|} \tag{7}$$

Regarding the *EE Precision* and *EE Recall*, we firstly calculate these two measures for each document $d \in D$ as follows:

$$\text{EE Precision}_d = \frac{|G_d^{EE} \cap A_d^{EE}|}{|A_d^{EE}|} \tag{8}$$

$$\text{EE Recall}_d = \frac{|G_d^{EE} \cap A_d^{EE}|}{|G_d^{EE}|} \tag{9}$$

Based on Eq. 8 and Eq. 9, the final *EE Precision* and *EE Recall* are averaged over all documents in D. The *EE F1* is the harmonic mean of *EE Precision* and *EE Recall*, calculated per document then averaged.

3.2 Evaluation Results

We evaluate our EED approach on top of a EL system, denoted by **RW-EE**$_{our}$, where we adopt a probabilistic EL model based on random walks [10], denoted by **RW**, which generates mention-entity mappings with their probabilities as a direct confidence measure. We compare our solution with two state-of-the-art

	EL Methods			EED Methods		
Measure	**AIDA**$_{sim}$	**AIDA**$_{coh}$	**RW**	**AIDA-EE**$_{sim}$	**AIDA-EE**$_{coh}$	**RW-EE**$_{our}$
Mic. Acc.	0.7602	0.7581	0.7616	0.7611	0.7133	**0.7900**
Mac. Acc.	0.7340	0.7258	0.7522	0.7290	0.7040	**0.7709**
EE Prec.	0.7284	0.5349	0.4328	**0.9797**	0.9392	0.8847
EE Rec.	0.8909	**0.9092**	0.7111	0.7069	0.7172	0.7478
EE F1	0.6661	0.4980	0.4023	0.6892	0.6792	**0.6954**

Table 2. Evaluation results (with the best results in bold font).

EED approaches [3], denoted by **AIDA-EE**$_{sim}$ and **AIDA-EE**$_{coh}$, which are accordingly based on two variants of the AIDA EL system [4], denoted by **AIDA**$_{sim}$ and **AIDA**$_{coh}$ respectively, where the difference lies in using keyphrase-based similarity or graph link-coherence for disambiguation. As additional baselines, we also consider the traditional EL methods, i.e., **RW**, **AIDA**$_{sim}$ and **AIDA**$_{coh}$, which all detect EE based on a threshold of the output confidence score.

Similar to [3], we estimate the parameters for all methods using the set of 150 documents from 2010-10-01 and based on that, the experiments are run on the 150 documents from 2010-11-01. All the methods use the same Stanford NER Tagger [7] for entity recognition, such that our comparison can focus on the ability of different methods to distinguish between existing and emerging entities, not the ability to recognize mentions in the input text.

The evaluation results in Table 2 clearly show that our approach **RW-EE**$_{our}$ achieves the best result in terms of EE F1. It is observed that the traditional EL methods (i.e., **AIDA**$_{sim}$, **AIDA**$_{coh}$ and **RW**) yield relatively high EE recall but low EE precision. This is because they determine EE only based on the absence of indication for KB entities such that a mention will be simply considered as an EE if there are no enough evidences for existing entities that can be extracted from the KB. Instead, the EED approaches (i.e., **AIDA-EE**$_{sim}$, **AIDA-EE**$_{coh}$ and **RW-EE**$_{our}$) detect EE by leveraging its direct positive indication harvested from external sources, where **RW-EE**$_{our}$ achieves an optimal trade off between EE precision and recall, which results in a better EE F1.

Furthermore, the results in Table 2 also show that **RW-EE**$_{our}$ outperforms all the competitors in terms of Micro Accuracy and Macro Accuracy in the full EL setting (w.r.t. both KB entities and EEs). While **RW-EE**$_{our}$ improves the performance of **RW** for the general EL, **AIDA-EE**$_{sim}$ and **AIDA-EE**$_{coh}$ yield degraded EL performance compared with **AIDA**$_{sim}$ and **AIDA**$_{coh}$ in some cases, such as Micro Accuracy for **AIDA-EE**$_{coh}$ and Macro Accuracy for both **AIDA-EE**$_{sim}$ and **AIDA-EE**$_{coh}$. This is due to the fact that for each mention **AIDA-EE**$_{sim}$ and **AIDA-EE**$_{coh}$ add an additional EE candidate for disambiguation and feed the expanded set of candidate entities back to the prior EL methods, i.e., **AIDA**$_{sim}$ and **AIDA**$_{coh}$, which would potentially introduce noise and result in degraded EL decisions for KB entities. In contrast, **RW-EE**$_{our}$ uses a prior EL method (i.e., **RW**) as a black box to generate the EE candidates for further EE detection, such that it does not affect existing EL decisions for KB entities yielded by the prior EL method.

4 Conclusions

In this paper, we aimed to address the challenge of discovering EEs in text by discriminating them against existing entities in the KB. In order to resolve the problems of existing methods, we devised a new EE detector for each mention as an EE candidate by comparing its KB entity context collected from Wikipedia and its actual entity context harvested from online Web sources based on the context representation learned as topic distributions. Our experiments show the superior quality of our solution in terms of higher F1 measure in detecting EEs compared with the state-of-the-art methods. More importantly, our approach considerably outperforms the existing methods in the full EL setting, where the measures of Micro Accuracy and Macro Accuracy w.r.t. both KB entities and EEs are considered. As future work, we would like to develop methods and tools to add the detected EEs with a canonicalized representation into the KB to improve its up-to-dateness and completeness.

References

1. Färber, M., Rettinger, A., Asmar, B.E.: On emerging entity detection. In: EKAW. (2016) 223–238
2. Fetahu, B., Anand, A., Anand, A.: How much is wikipedia lagging behind news? In: WebSci. (2015) 28:1–28:9
3. Hoffart, J., Altun, Y., Weikum, G.: Discovering emerging entities with ambiguous names. In: WWW. (2014) 385–396
4. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: EMNLP. (2011) 782–792
5. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: Dbpedia: A nucleus for a web of open data. In: ISWC. (2007) 722–735
6. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW. (2007) 697–706
7. Finkel, J.R., Grenager, T., Manning, C.D.: Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL. (2005) 363–370
8. Kulkarni, S., Singh, A., Ramakrishnan, G., Chakrabarti, S.: Collective annotation of wikipedia entities in web text. In: KDD. (2009) 457–466
9. Ratinov, L., Roth, D., Downey, D., Anderson, M.: Local and global algorithms for disambiguation to wikipedia. In: ACL. (2011) 1375–1384
10. Han, X., Sun, L., Zhao, J.: Collective entity linking in web text: a graph-based method. In: SIGIR. (2011) 765–774
11. Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., Manning, C.D.: A multi-pass sieve for coreference resolution. In: EMNLP. (2010) 492–501
12. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: HLT-NAACL. (2003)
13. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
14. Griffiths, T.L., Steyvers, M.: Finding scientific topics. PNAS **101**(suppl 1) (2004) 5228–5235
15. Parker, R.: English gigaword fifth edition. Technical report (2011)